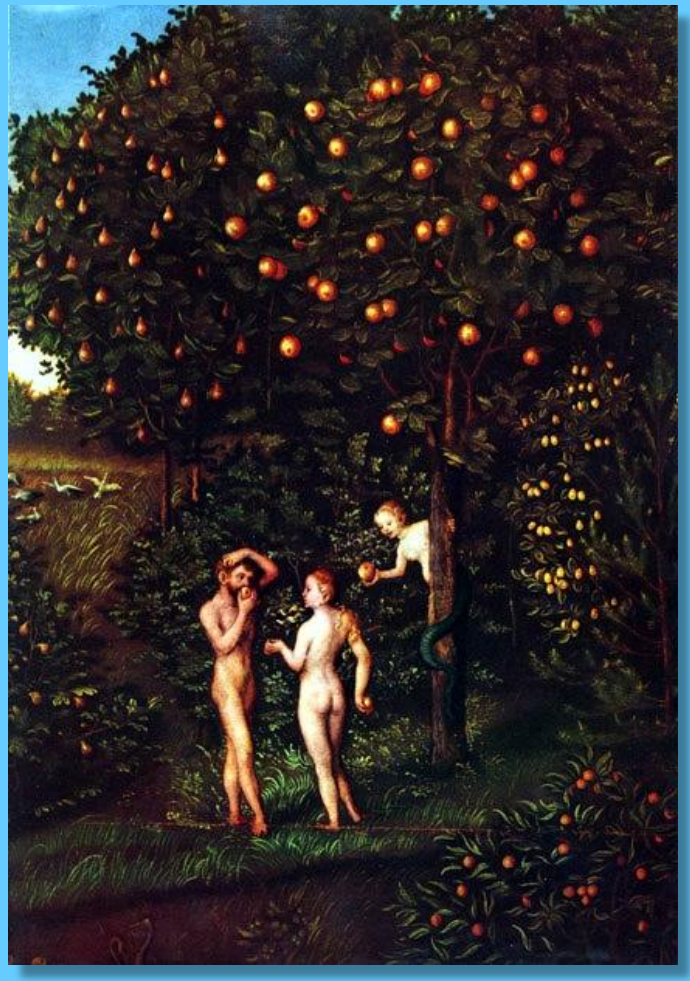
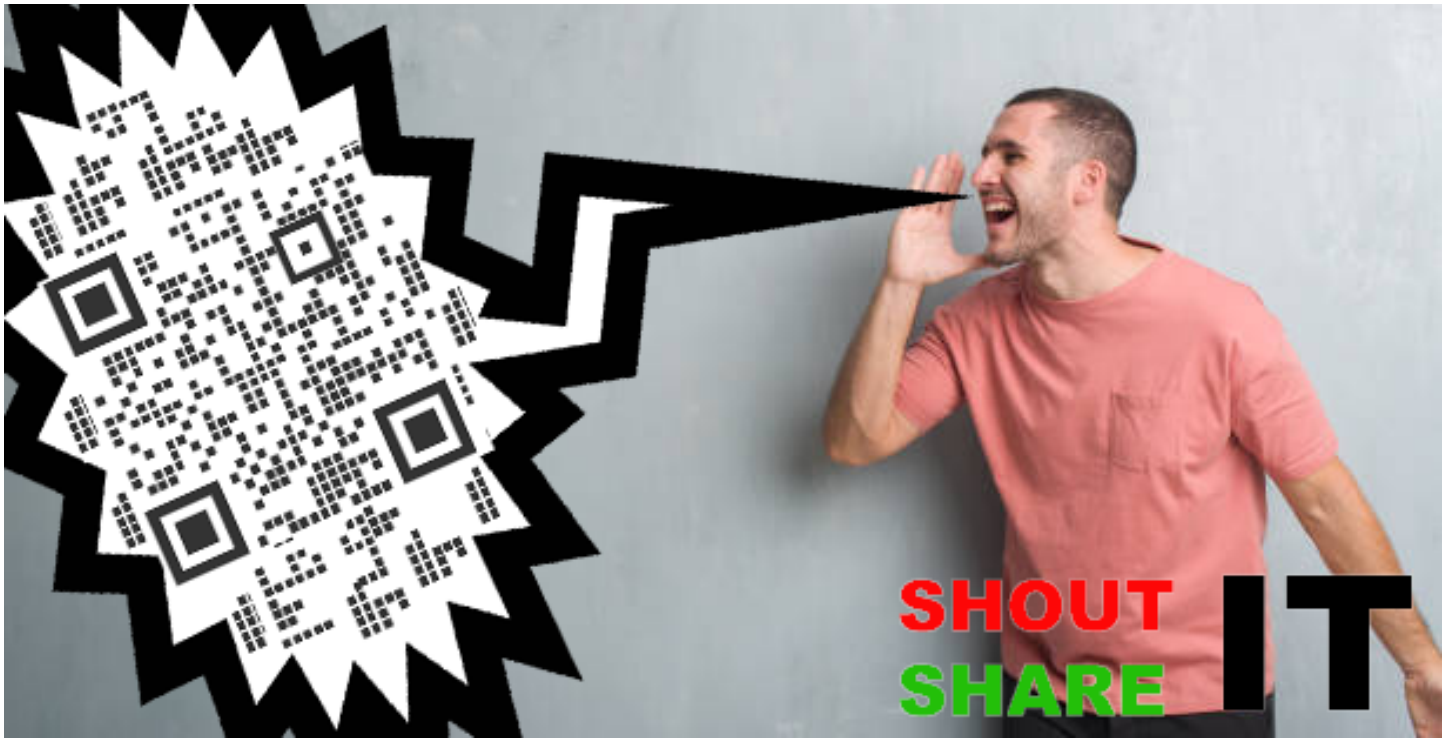
Statistics: The Z score and the normal distribution

- ✓ Click “slide show” to start this presentation as a show.
- ✓ **Remember: focus & think about each point; do not just passively click.**



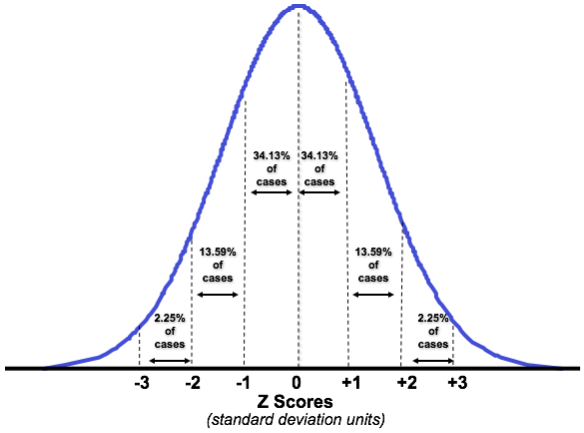
Cranach, *Tree of Knowledge* [of Good and Evil] (1472)

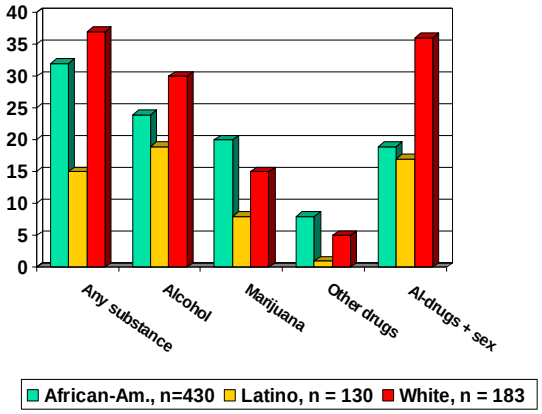
© Dr. David J. McKirnan, 2014
The University of Illinois Chicago
McKirnanUIC@gmail.com
Do not use or reproduce without permission


SHOUT
SHARE
IT

1. Introduction to statistics & number scales
2. The Z score and the normal distribution
3. The logic of research; Plato's Allegory of the Cave
4. Testing hypotheses: The critical ratio
5. Calculating a t score
6. Testing t: The Central Limit Theorem
7. Correlations: Measures of association

You are here



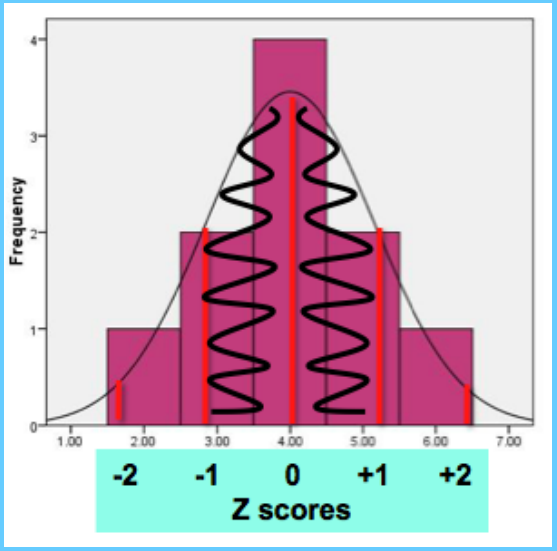


© Dr. David J. McKirnan, 2014
The University of Illinois Chicago
McKirnanUIC@gmail.com
Do not use or reproduce without permission

This module covers two topics:

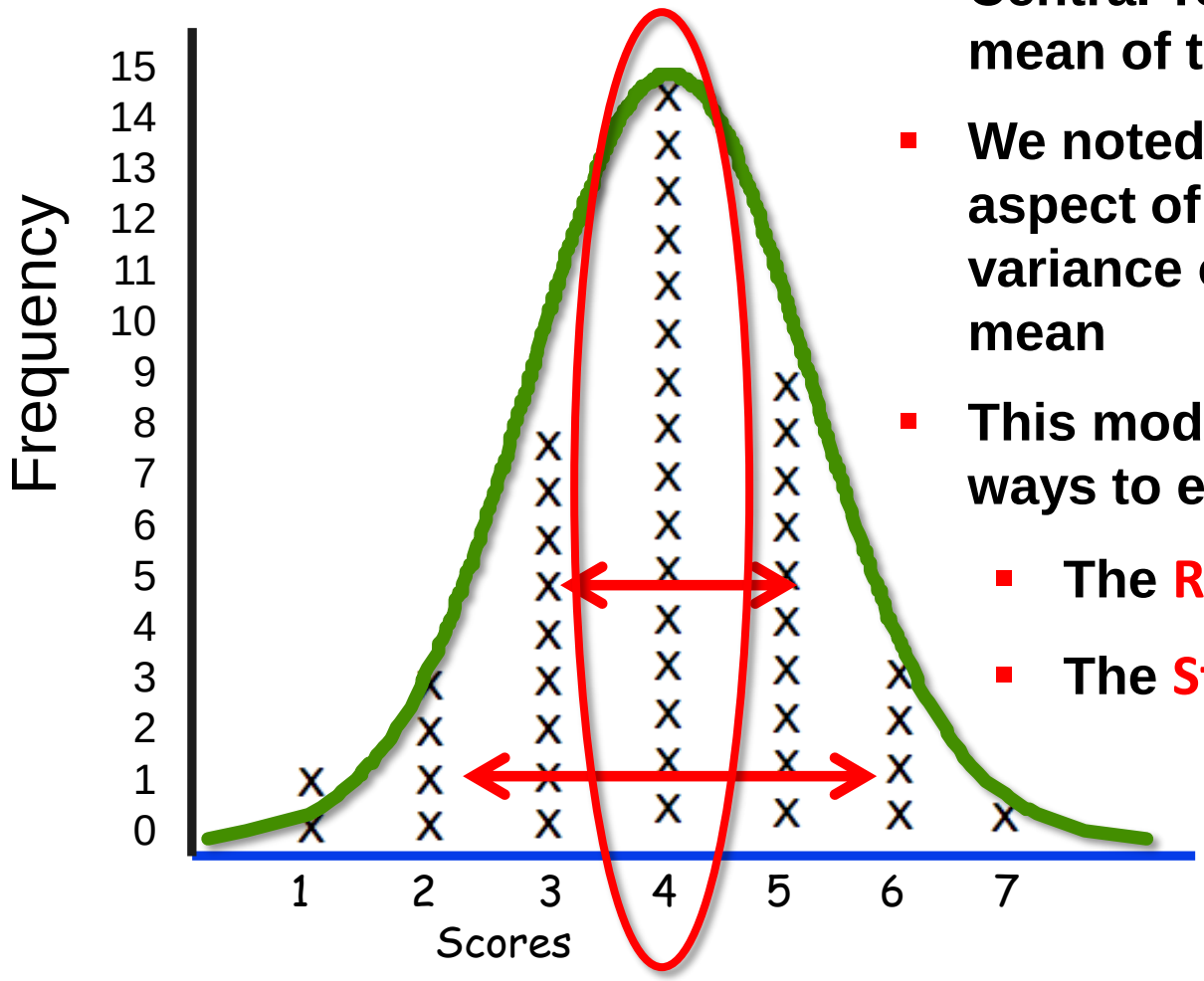
- Variance: The **Standard Deviation**

- The **Z score** and the normal distribution



© Dr. David J. McKirnan, 2014
The University of Illinois Chicago
McKirnanUIC@gmail.com
Do not use or reproduce without permission

Variance



In module 1 we discussed

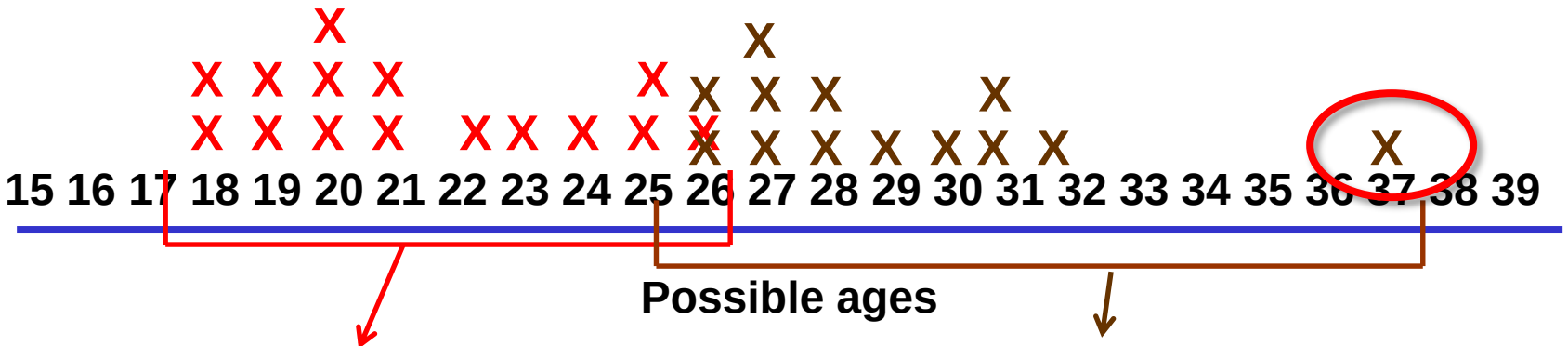
- Distributions of scores
- Central Tendency, such as the mean of the scores
- We noted that a 2nd important aspect of a distribution is the variance of scores around the mean
- This module will describe two ways to express variance:
  - The Range
  - The Standard Deviation

1. The Range of the highest to the lowest score.

- The range is easy to compute and understand, but can be misleading where there is a lot of variance in scores
- Imagine we are comparing ages of male and female samples

Ages of males: 18, 25, 20, 21, 20, 23, 24, 26,18, 25, 20, 19, 19.

Ages of women: 26, 27, 27, 31, 32, 28, 31, 29, 30, 27, 26, 37, 28



Scores (ages) in the male sample range from 18 to 26, range (26-18) = 8.

Scores in the female sample range from 26 to 37, range (37-26) = 11.

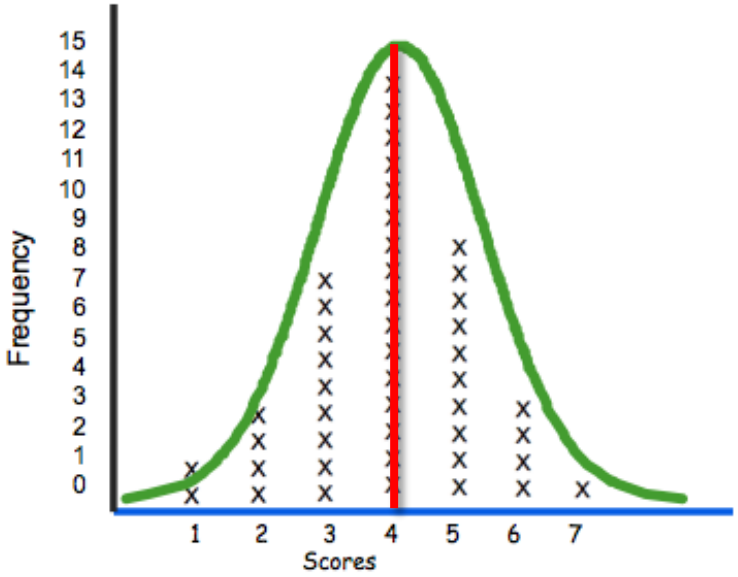
→ Note: most female scores are in a smaller range than the men: the range is very sensitive to extreme values.

2. The Standard deviation of scores around the Mean (*S*)

- **Similar to the "*average*" amount each score deviates from the *M* of the sample.**
- **"Standardizes" scores to a normal curve, allowing basic statistics to be used.**
- **More accurate & detailed than range:**
  - ***A few extremely high or low scores ("outliers") may make the range inaccurate***
  - ***S assesses the deviation of all scores in the sample from the mean***

Standard deviation; Basic Steps

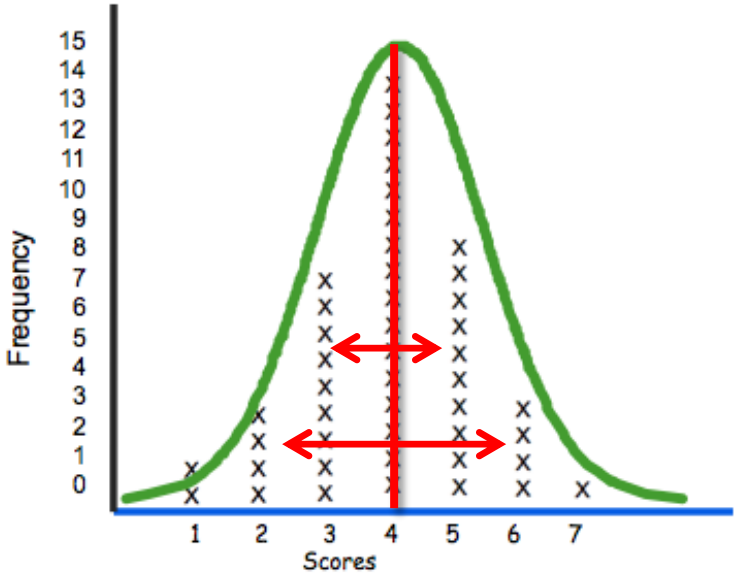
1. **Calculate the Mean score**



Use the Mean [*M*] to assess the Central Tendency of the scores in the sample.

Standard deviation; Basic Steps

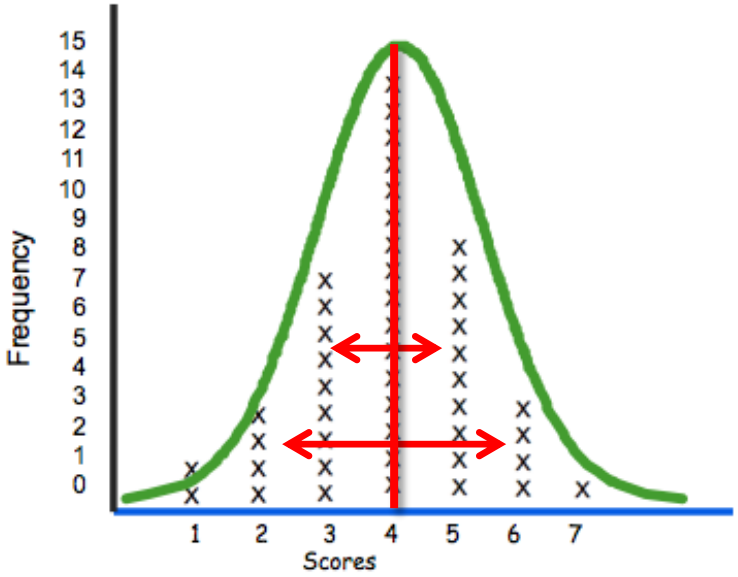
1. **Calculate the Mean score**
2. **Express each score as a deviation from the M**



This provides the basic index of how much the scores vary around the Mean

Standard deviation; Basic Steps

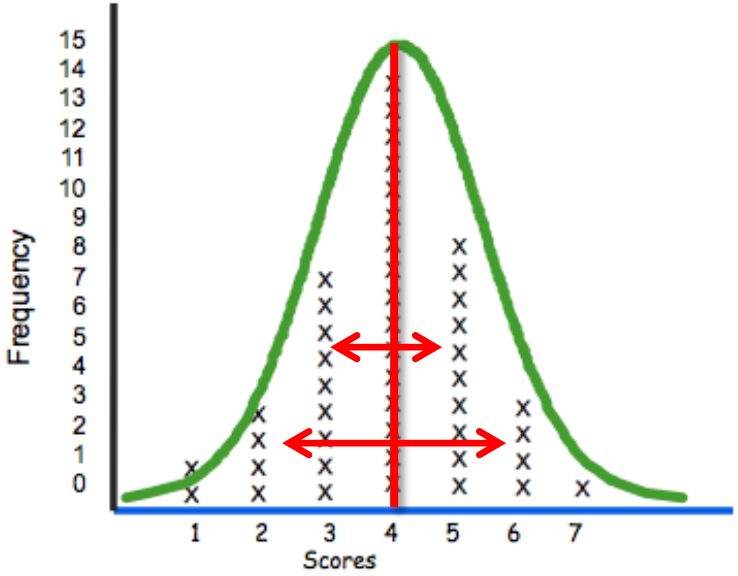
1. Calculate the Mean score
2. Express each score as a deviation from the M
3. Square each deviation score



Squaring the deviation scores keeps them from all just adding up to 0.

Standard deviation; Basic Steps

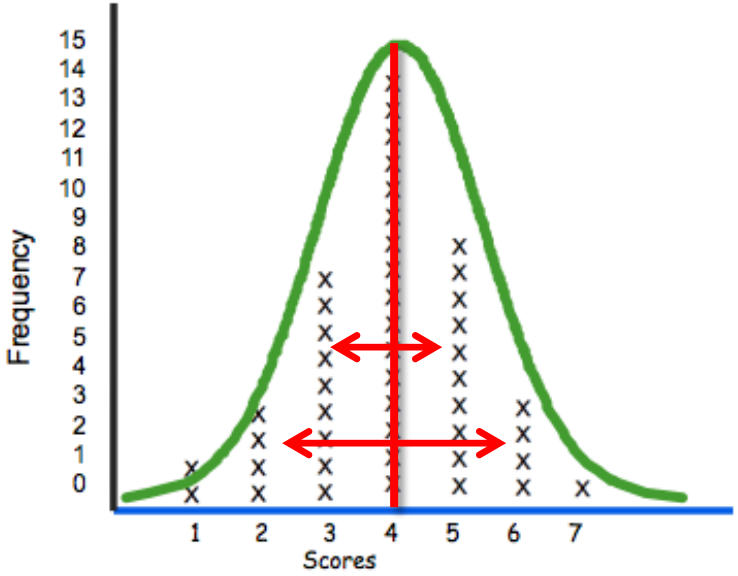
1. Calculate the Mean score
2. Express each score as a deviation from the M
3. Square each deviation score
4. Sum the squared deviation scores



Sum the squared deviations to calculate the total amount the scores vary – known as the “sum of squares”.

Standard deviation; Basic Steps

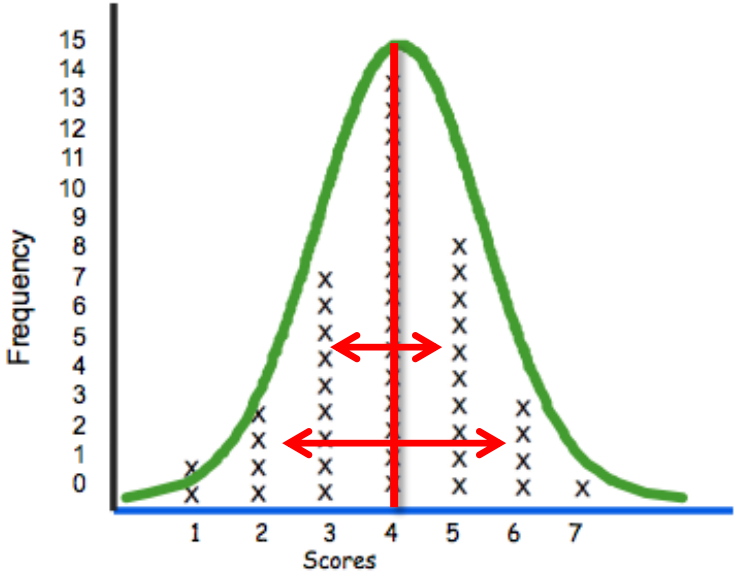
1. Calculate the Mean score
2. Express each score as a deviation from the M
3. Square each deviation score
4. Sum the squared deviation scores
5. Divide by the degrees of freedom



Divide by the number of scores that can vary – the degrees of freedom [*df*] (see below).

Standard deviation; Basic Steps

1. Calculate the Mean score
2. Express each score as a deviation from the M
3. Square each deviation score
4. Sum the squared deviation scores
5. Divide by the degrees of freedom
6. Take the square root of the result.



Since we squared the original deviation scores, take the square root of this result to put the numbers back into the original scale

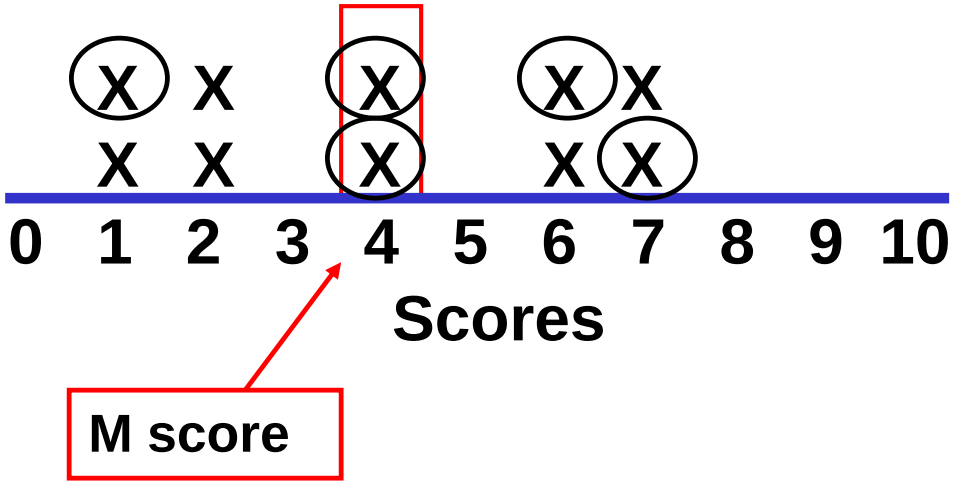
1. Take a set of scores:

X = 7, 6, 2, 1, 4, 1, 7, 4, 2, 6.

2. Calculate the Mean:

$$M = \sum X / n = 40 / 10 = 4$$

3. Express each score as a deviation from the M; ($X - M$).



Deviation Scores:

0, 0, +2, -3, +3…

- The Σ of deviations ($X - M$) *must* = 0
- **Standard Deviation (S)** adjusts by squaring each deviation $(X - M)^2$ and then summing; $\Sigma (X - M)^2$

Standard Deviation & Formulas

X	**Score** on one variable for one participant
n	**Number** of scores in the sample
Σ	**Sum** of a set of scores
M	**Mean**; sum of scores divided by n of scores: $\sum x / n$

X - M **Deviation** of one score from the mean

$(X - M)^2$ **Squared deviation** of score from mean

SS **Sum** of **Squared** deviations from the mean.

$$\Sigma (X - M)^2$$

Degrees of freedom (*df*): the number of scores that can vary…

- Assume you know that the sum of a set of 5 scores is 20 (n = 5, Σ = 20).
- If you know the first 3 scores, scores 4 & 5 could be almost any combination..
- If you know the first 4 scores, the 5th score is determined
- …here it *must* be 3

Score
$X_1 = 6$
$X_2 = 4$
$X_3 = 2$
$X_4 = 5$
$X_5 = 3$
$\Sigma = 20$

- ✓ With 5 scores (n = 5), we have 4 degrees of freedom (df = 4)
- ✓ Degrees of freedom typically = n - 1

Degrees of freedom (*df*): the number of scores that can vary…

Technically, df is the number of independent observations in our data, minus the number of parameters to be estimated.

Here we have one group, and n = 10;

- ✓ we are estimating 1 parameter, the group mean
- ✓ so df = n – 1 (10 - 1 = 9).

Say these data were for men and women:

What are the df here?

N = 10, but we are estimating two parameters: Means for **two** groups, so df is not n-1, rather it is:

$$N_{women} - 1 \quad + \quad N_{men} - 1$$

(10 observations minus 2 parameters = 8.)

Women		Men	
X1 =	6	X_6 =	5
X_2 =	4	X_7 =	2
X3 =	2	X_8 =	7
X4 =	5	X_9 =	5
X5 =	3	X_{10} =	2

Scores
Scores

X_1 = 6
X_2 = 4
X_3 = 2
X_4 = 5
X_5 = 3
X_6 = 5
X_7 = 2
X_8 = 7
X_9 = 5
X_{10} = 2

Standard Deviation & Formulas

X	**Score** on one variable for one participant
n	**Number** of scores in the sample
Σ	**Sum** of a set of scores
M or $\overline{X}$	**Mean**; sum of scores divided by n of scores: $\sum x / n$
X - M	**deviation** of one score from the mean
$(X - M)^2$	**squared deviation** of score from mean
SS	**sum of squared** deviations from the mean: $\Sigma (X - M)^2$
df	**degrees of freedom**; # of scores that are free to vary; $n - 1$

Variance example

How many hours per day do you spend studying research methods?

Name	# hours (score, or 'X')
Bill	7
Joe Bob	6
Sally	2
Eloise	1
William	4
Robert	1
Barak	7
Hank	4
Glenn	2
Mary Louise	6

What is the average?

Mean: ΣX / n = 40/10 = 4

How much variance is there?

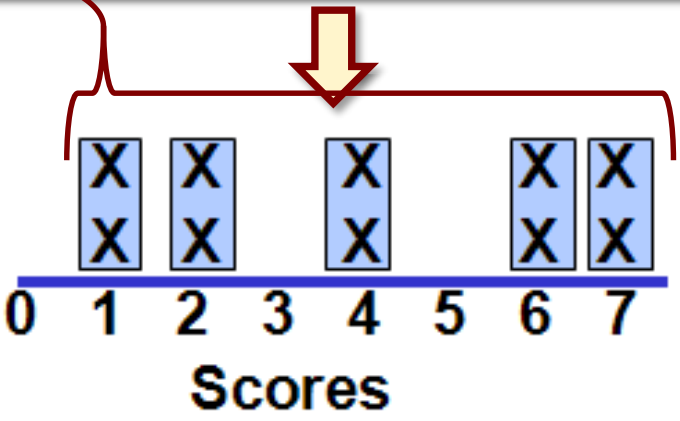
How consistent are these scores?

How much do these scores vary?

- This is a “flat”, wide distribution; lots of variance

Name	# hours (score, or ‘X’)
Bill	7
Joe Bob	6
Sally	2
Eloise	1
William	4
Robert	1
Barak	7
Hank	4
Glenn	2
Mary Louise	6

- The Range = 6.
- Calculate the Standard Deviation (S) to better show overall variance.
- In this example S = 2.4
- How did we compute that?



Calculating the standard deviation

X	*M*	*X - M*	$(X - M)^2$
7	4	*3*	9
6	4	*2*	4
2	4	*-2*	4
1	4	*-3*	9
4	4	*0*	0
1	4	*-3*	9
7	4	*3*	9
4	4	*0*	0
2	4	*-2*	4
6	4	*2*	4
n = 10			
Σ= 40		**Σ = *0***	**Σ = 52**
M = 40/10 = 4			

1. Calculate the Mean score:
   ΣX / n = 40 / 10 = 4

2. Calculate how much each score deviates from the M

→ The Sum of the simple deviations: Σ (X – M) will always = 0

→ Square the deviations to create + values: Σ Squares = $\Sigma(X - M)^2 = 52$

3. Degrees of freedom:
   df = n - 1 = 9

4. Now calculate the variance (S^2):

- Take the sum of the squared deviations: $\Sigma (X-M)^2$
- Divide by the df

Calculating the standard deviation

X	M	X - M	$(X - M)^2$
7	4	3	9
6	4	2	4
n = 10			
Σ= 40		Σ = 0	Σ = 52
M = 40/10 = 4			

- **We squared all the deviation scores to make them positive numbers.**
- **To get back to the original scale we take the square root of the variance.**
- **The Standard Deviation (S):**

$$S = \sqrt{S^2}_{(variance)}$$

$$S = \sqrt{5.8} = 2.4$$

1. Calculate the Mean score:
   ΣX / n = 40 / 10 = 4

2. Calculate how much each score deviates from the M

→ The Sum of the simple deviations: Σ (X – M) will always = 0

→ Square the deviations to create + values: Σ Squares = $\Sigma(X - M)^2 = 52$

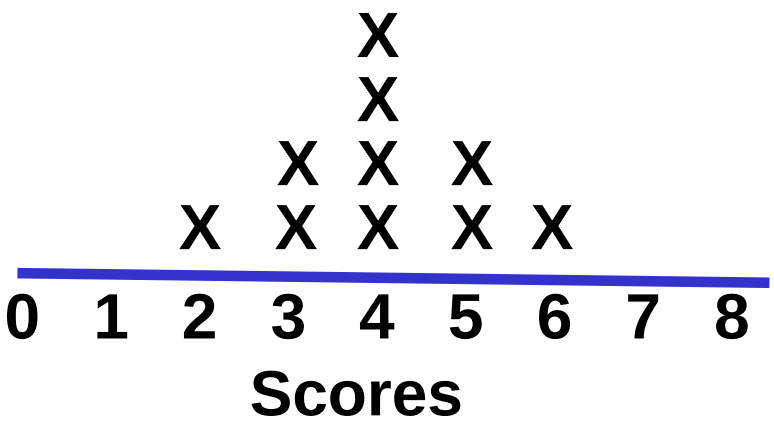
3. Degrees of freedom:
   df = n - 1 = 9

4. Now calculate the variance (S^2):

$$S^2 = \frac{\sum (X - M)^2}{df} = \frac{52}{9} = 5.8$$

Scores with less variance

How much do these scores vary?



- **This is a more normal, "tighter" distribution**
- **The Range = 4** (6-2)**.**
- **The Standard Deviation = 1.15** (the standard deviation is lower, reflecting the lower variance in this distribution…)

In a distribution with scores closer to the *M* the Standard Deviation goes down…

X	*M*	*X - M*	$(X - M)^2$
4	4	*0*	0
3	4	*-1*	1
5	4	*1*	1
5	4	*-1*	1
4	4	*0*	0
2	4	*-2*	4
4	4	*0*	0
4	4	*0*	0
3	4	*-1*	1
6	4	*2*	4
n = *10* Σ = *40*		Σ = *0*	Σ = *12*

Variance formula:

1. Mean ΣX / n = 40/10 = 4

2. Deviation scores:
Σ of Squares: $\Sigma (X - M)^2 = 12$

3. Degrees of freedom:
df = n - 1 = 9

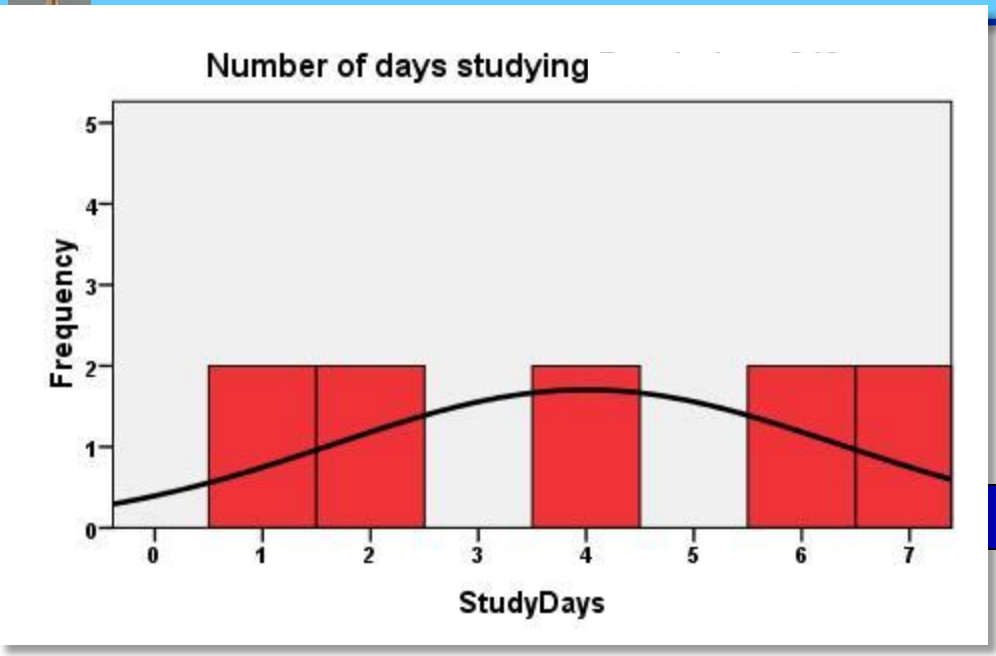
4. Variance:

$$S^2 = \frac{\sum (X-M)^2}{df} = \frac{12}{9} = 1.33$$

5. Standard Deviation:

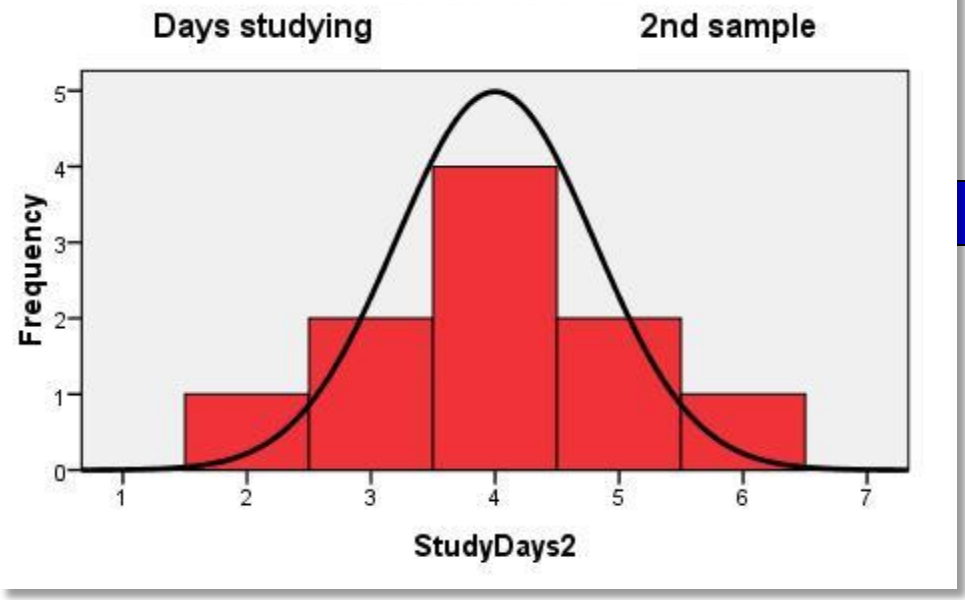
$$S = \sqrt{S^2}\,(\text{variance}) = 1.15$$

Differing variances



The data sets have the same M, but differ in how widely their scores vary (their variance).

- ✓ M = 4
- ✓ High variance;
- ✓ S = 2.4



- ✓ M = 4,
- ✓ Less variance;
- ✓ S = 1.15

X	**Score** on one variable for one participant
n	**Number** of scores in the sample
Σ	**Sum** of a set of scores
M	**Mean**; sum of scores divided by n of scores: $\sum x / n$
X - M	**deviation** of one score from the mean
$(X - M)^2$	**squared deviation** of score from mean
SS	**sum of squared** deviations from the mean: $\Sigma (X - M)^2$
df	**degrees of freedom**; # of scores that are free to vary; $n - 1$

S^2 **Variance sum of squared deviations from *M* divided by degrees of freedom:**

$$SS/df = \sum (X - M)^2 / n - 1$$

S **Standard Deviation, square root of the variance:**

$$\sqrt{\sum (X - M)^2 / n - 1}$$

The number of scores that are free to vary in a given simple is called the…

A. Mean
B. Standard Deviation
C. Degrees of Freedom
D. Sum of Squares
E. Variance
F. Range

The number of scores that are free to vary in a given simple is called the…

A. Mean
B. Standard Deviation
C. Degrees of Freedom
D. Sum of Squares
E. Variance
F. Range

df **is typically calculated as n = 1.**

It reflects the degree of "flexibility" in a set of scores.

We use this in many calculations, including the Standard Deviation.

Both the range and the standard deviation are examples of this…

A. Mean
B. Standard Deviation
C. Degrees of Freedom
D. Sum of Squares
E. Variance
F. Range

Both the range and the standard deviation are examples of this…

A. Mean
B. Standard Deviation
C. Degrees of Freedom
D. Sum of Squares
E. Variance
F. Range

"Variance" has two meanings in statistics:

- ✓ **The general concept of scores differing from each other in a sample**
- ✓ **A statistical formula, part of the calculation of the Standard Deviation.**

Represents a sort of “average” amount that scores vary around the M…

A. Mean
B. Standard Deviation
C. Degrees of Freedom
D. Sum of Squares
E. Variance
F. Range

Represents a sort of "average" amount that scores vary around the M…

A. Mean
B. Standard Deviation
C. Degrees of Freedom
D. Sum of Squares
E. Variance
F. Range

The Standard Deviation (S) is sensitive to how far all the scores in the distribution are from the mean.

If we add up (or take the average of) how far each individual score is from the M, we will get…

A. Z
B. 1
C. M / n-1
D. 0
E. Variance
F. Range

If we add up (or take the average of) **how far each individual score is from the M, we will get…**

A. Z
B. 1
C. M / n-1
D. 0
E. Variance
F. Range

- ✓ **M is in the center of the distribution,**
- ✓ **Any score a given amount above it *must* correspond to a score equally below it.**
- ✓ **So, adding deviation scores [Σ (X - M)] always = 0.**

Summary

Central tendency:

- For normal distributions we use the Mean [*M*]; $M = \sum x / n$

Variance:

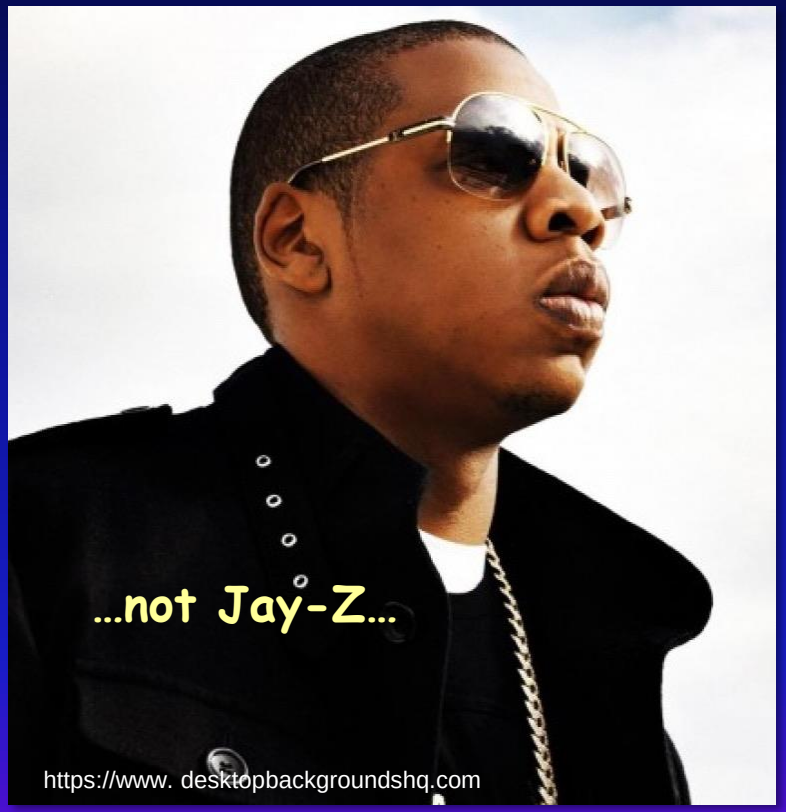
- The range expresses the span of the highest to lowest score
  - ✓ Easy and comprehensible description of data
  - ✓ Very sensitive to extreme values (“outliers”)
- Standard Deviation [*S*] of cases around the *M* is the most common measure of variance

$$\sqrt{\sum (X - M)^2 / n - 1}$$

  - ✓ Includes all the scores in the distribution
  - ✓ Basic to statistical testing; reflects the “error” in our measurement.

Variance: The **Standard Deviation**

- ✓ The **Z score** and the normal distribution



How do we characterize how high or low one score is?

- ✓ On an attitude scale…
- ✓ The Dependent Variable in an experiment…
- ✓ Elapsed time…

We use three pieces of information:

- The individual Score → [X]
- The Central Tendency of all the scores in the sample; → Mean [M]
- The Variance of the scores around the *M*: → Standard Deviation [S]

- How do we combine these into a single **metric** (mathematical description) to characterize a score?
- **Z score:**

$$\frac{\text{How far is this individual score from the M?}}{\text{How much variance is there around the M?}} = X - M / S$$

$$Z: \frac{\text{Distance of a score from M}}{\text{Standard Deviation of scores around the M}} = \frac{X - M}{S}$$

Z expresses the strength of a score relative to all other scores in the sample.

Rather than using literal scale value

e.g., elapsed time to task completion, a rating scale value…

or how far the score is above / below the M

Z expresses the score as:

$$\frac{\text{How far the score varies from the M}}{\text{The amount of variance in all the scores}}$$

…or, the % of scores it is above / below in the distribution.

This allows us to use the Normal Distribution to interpret the score.

Properties of the normal distribution

- The normal distribution is a **hypothetical distribution** of cases in a sample
- It is segmented into **standard deviation units**, denoted by Z
- Each standard deviation unit (Z) has a fixed % of cases above or below it.
- A given Z score,

  tells you the % of scores in the sample lower than yours

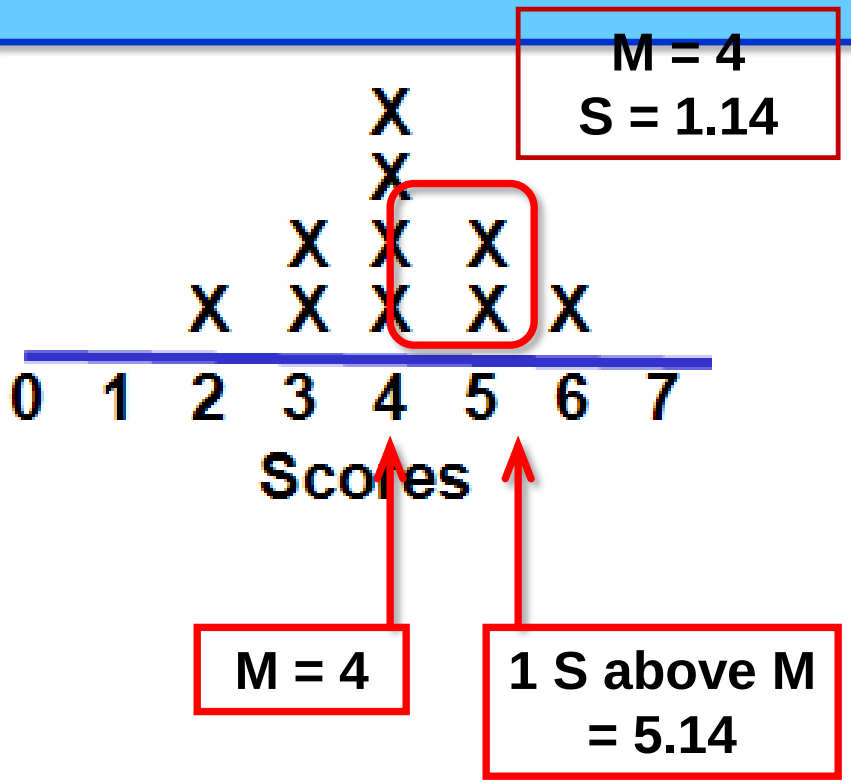
  e.g., Z = 1,

  84% of scores are below Z = 1.

- We use **Z scores** & associated % of the normal distribution to make statistical decisions about whether a score might occur by chance.

In this distribution…

- ✓ There are a specific % of cases between the **M** [4] and one standard deviation (**S**) above the mean



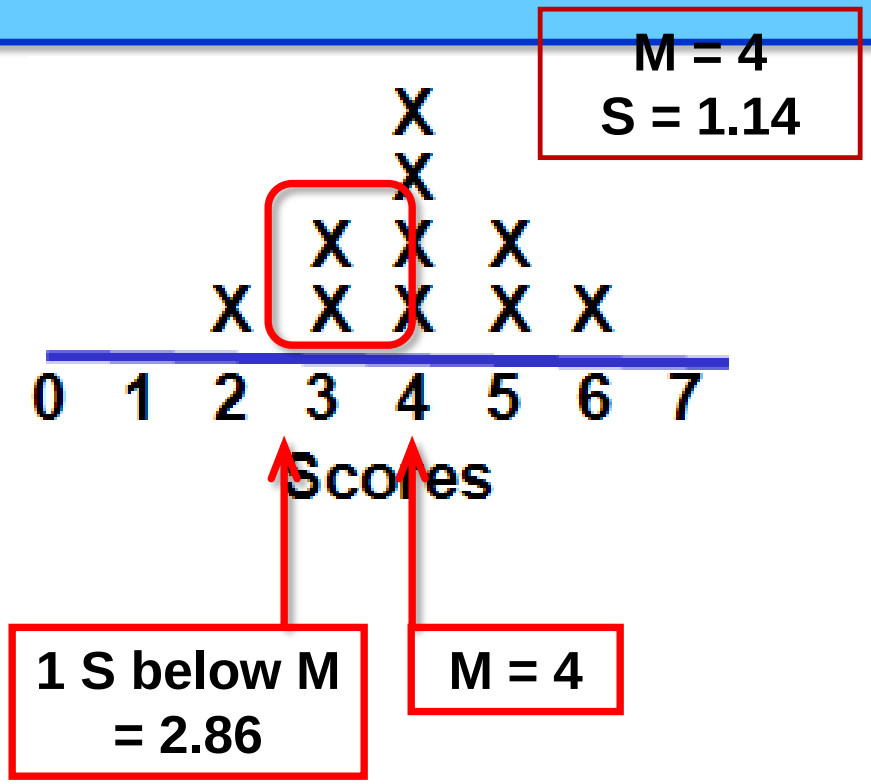
Hint:

- The Mean is 4
- The Standard Deviation is 1.14
- A score of 5.14 is 1 Standard Deviation above the Mean

In this distribution…

- ✓ There are the same % of cases between the **M** [4]

  and one standard deviation (**S**) BELOW the mean.

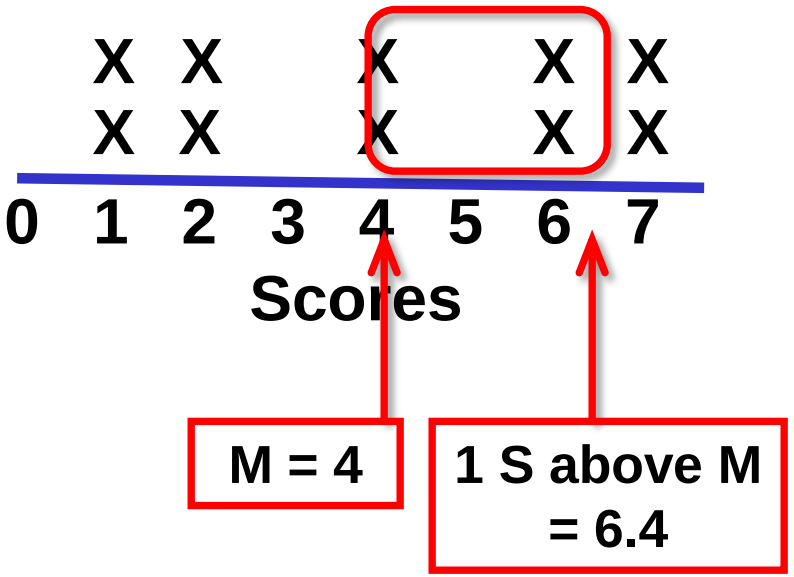


Hint: 4 (**M**) – 1.14 (**S**) = 2.86

M = 4
S = 2.4

This distribution…

- ✓ Has the exact same % of cases between the **M** [4] and one standard deviation (**S**) above the mean as the other distribution.

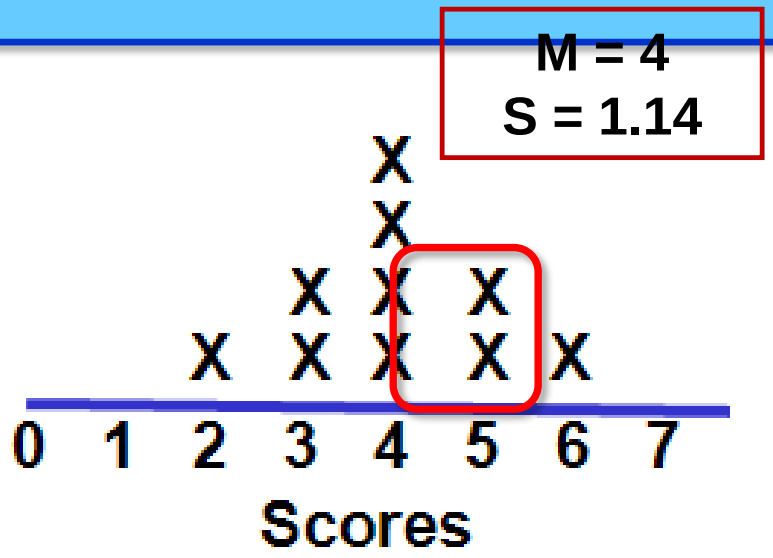


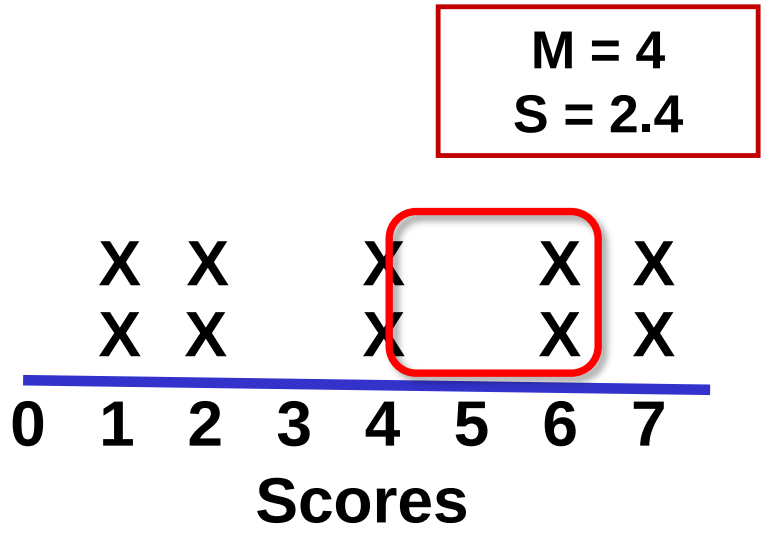
- ✓ This is because **S** is based on the distribution of cases in our particular sample.

Hint: 4 (**M**) + 2.4 (**S**) = 6.4

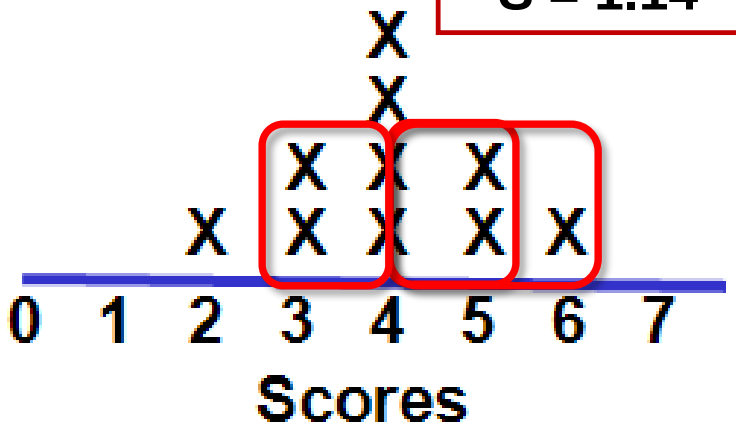
So…

- ✓ No matter what the sample is…
- ✓ …what the **M** is
- ✓ …or what the variance is in the distribution…
- ✓ One **S** above (or below) the **M** will always constitute the exact same % of cases.





M = 4
S = 1.14



- ✓ This allows us to ***segment*** a distribution into standard deviation units
  - One standard deviation above the **M** [4 → 5.14]
  - Two standard deviations above **M** [4 → 6.28]
  - One **S** below the **M** [4 → 2.86]
- ✓ Each segment represents a certain % of cases.
- ✓ These segments are denoted by Z scores

Z scores

$$Z = \frac{X - M}{S} = \frac{\text{Individual score} - \text{M for sample}}{\text{Standard deviation for sample}}$$

- Z describes how far a score is above or below the M in *standard deviation units* rather than raw scores.
  - "Adjusts" the score to be independent of the original scale.
  - We transform the original scale – inches, elapsed time, performance – into universal standard deviation units.
- Z allows us to use the general properties of the normal distribution to determine how much of the curve a score is above or below.

Standard Deviation & Formulas

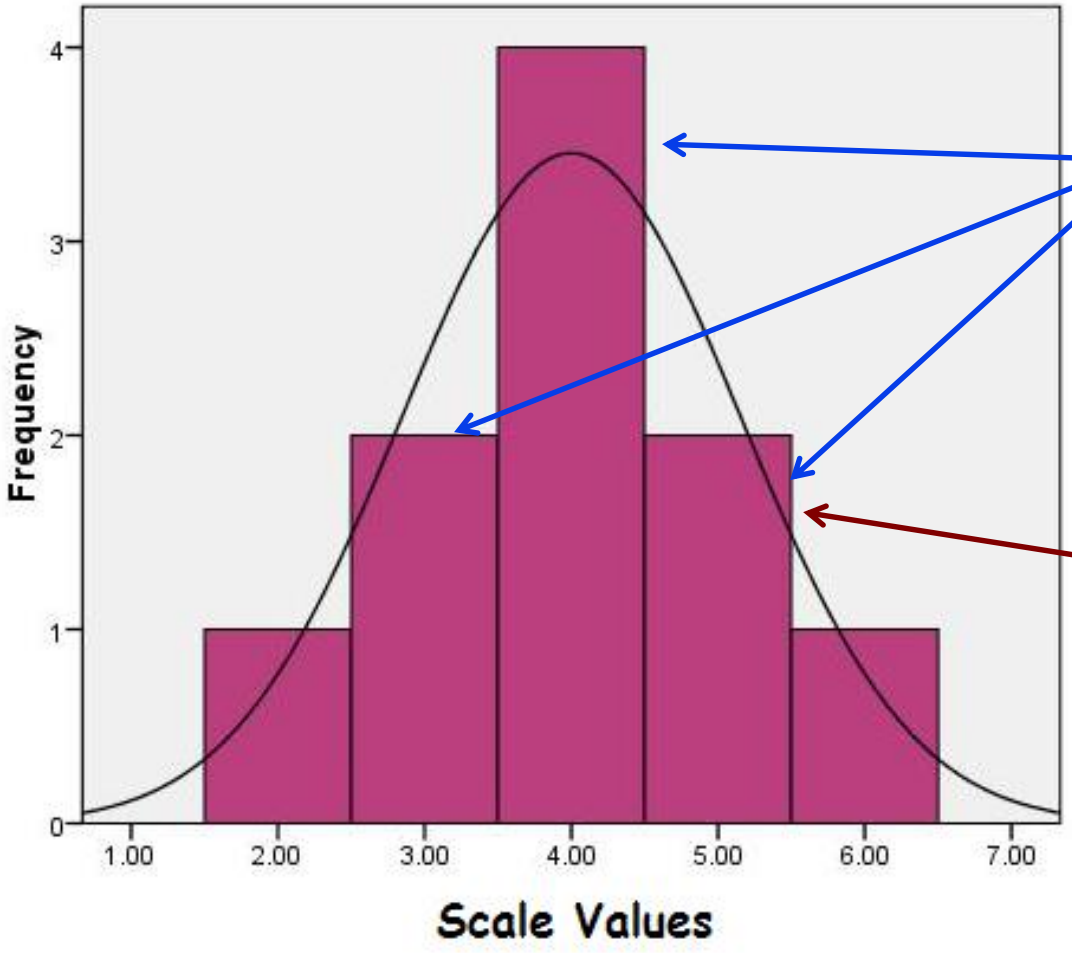
X	**Score** on one variable for one participant
n	**Number** of scores in the sample
Σ	**Sum** of a set of scores
M or $\overline{X}$	**Mean**; sum of scores divided by n of scores: $\Sigma x / n$
$X - M$	**deviation** of one score from the mean
$(X - M)^2$	**squared deviation** of score from mean
SS	**sum of squared** deviations from the mean: $\Sigma (X - M)^2$
df	**degrees of freedom**; # of scores that are free to vary; $n - 1$
S^2	**Variance** sum of squared deviations from M divided by degrees of freedom: $=$ $\Sigma (X - M)^2 / n - 1$
S	**Standard Deviation,** square root of the variance: $\sqrt{\Sigma (X - M)^2 / n - 1}$

Z score **# of standard deviation units:**

Difference between score & mean, divided by standard deviation

$$X - M / S$$

We use Z scores based on a hypothetical sampling distribution



Frequency distribution we observe in our sample

Hypothetical frequency distribution in the population if it had the same statistical characteristics as our sample

The Normal Distribution

34.13% of scores from Z = 0 to Z = +1 **and** from Z = 0 to Z = -1

13.59% of scores + 13.59% of scores

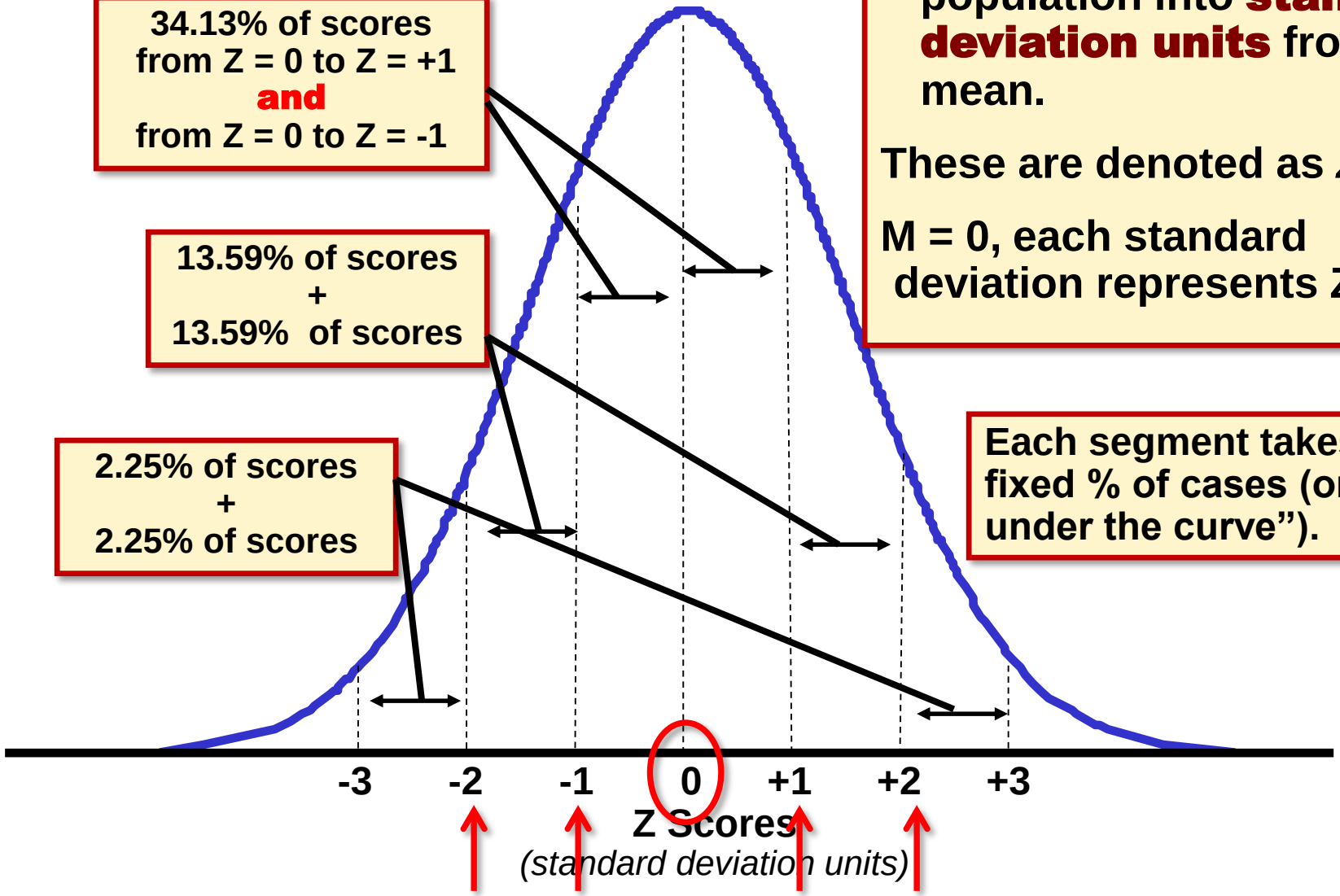
2.25% of scores + 2.25% of scores

We can segment the population into **standard deviation units** from the mean.

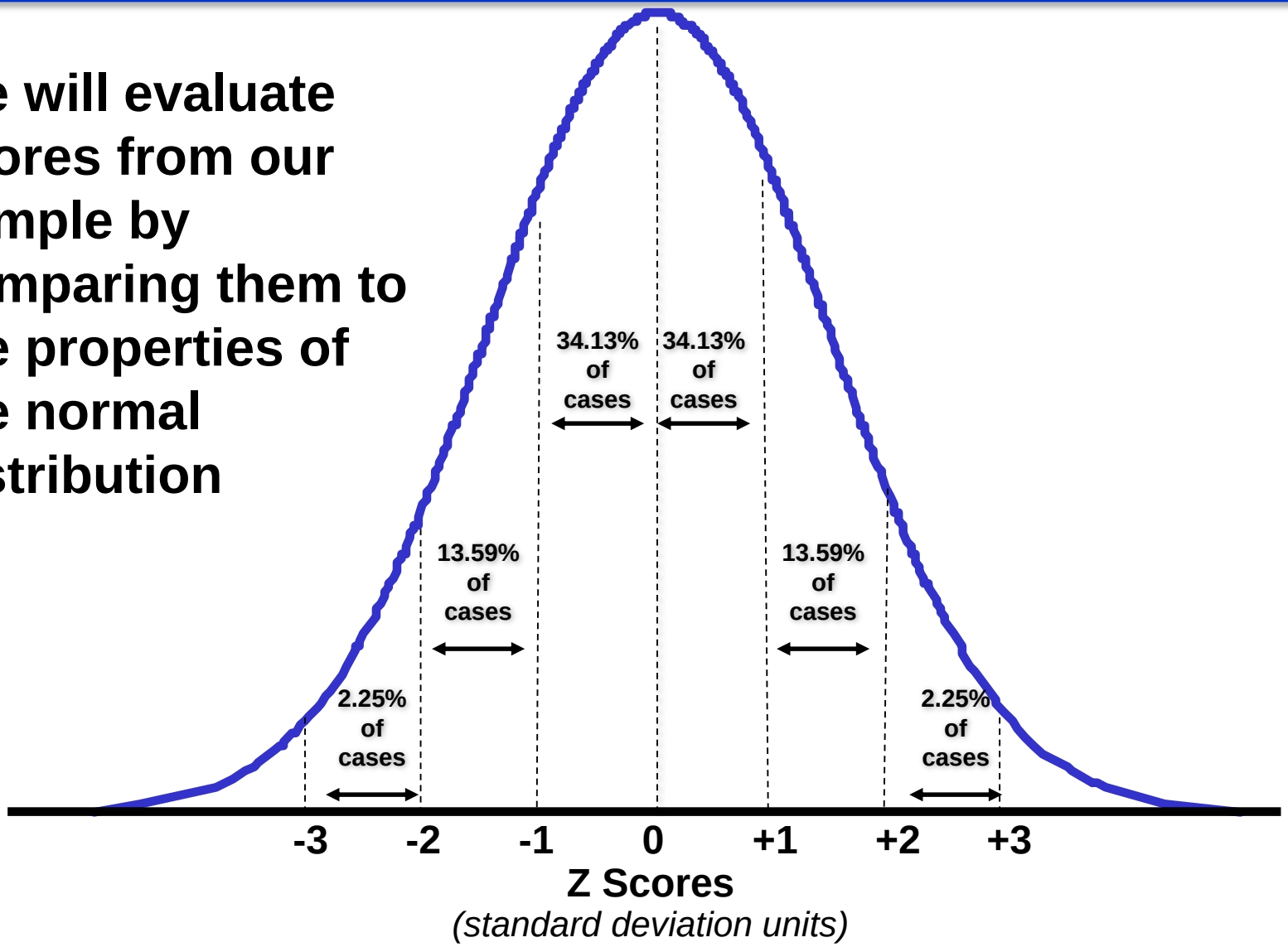
These are denoted as *Z*

M = 0, each standard deviation represents Z = 1

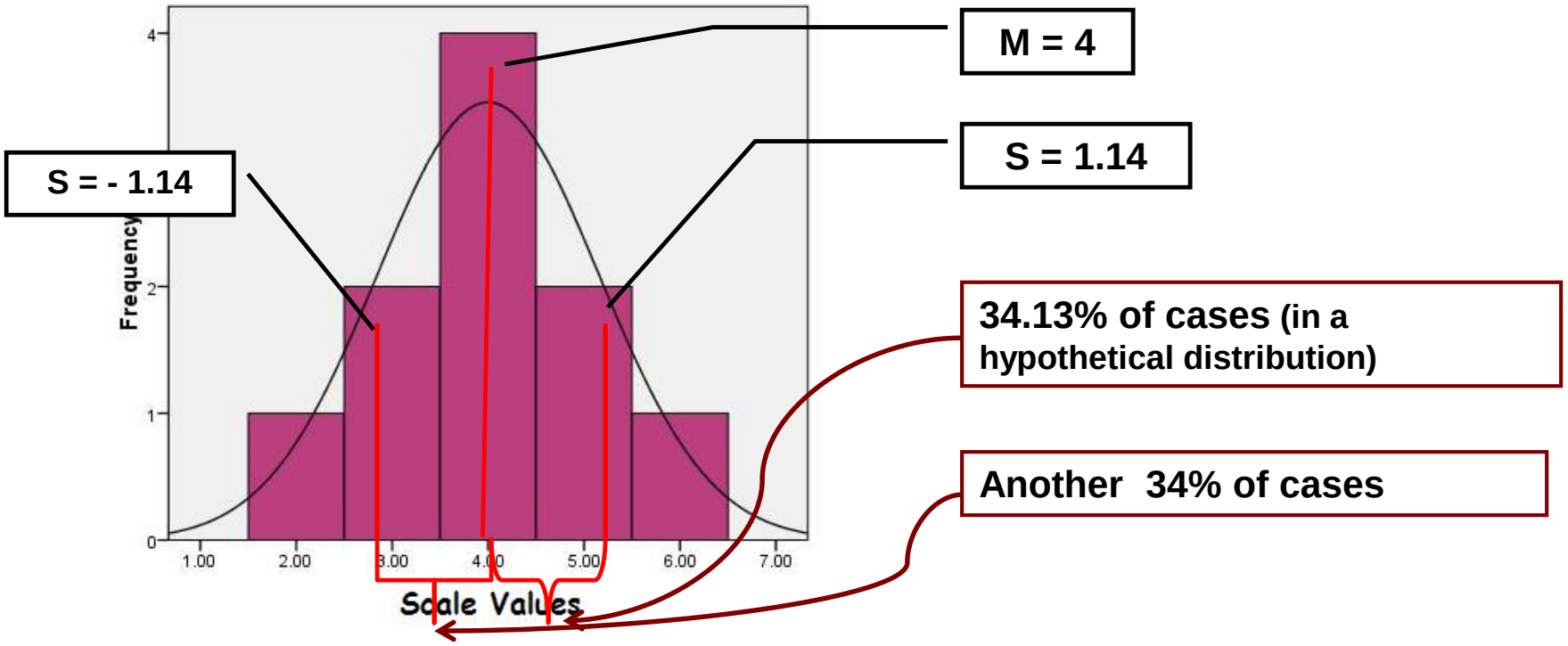
Each segment takes up a fixed % of cases (or "area under the curve").



We will evaluate scores from our sample by comparing them to the properties of the normal distribution



Standard deviations and distributions

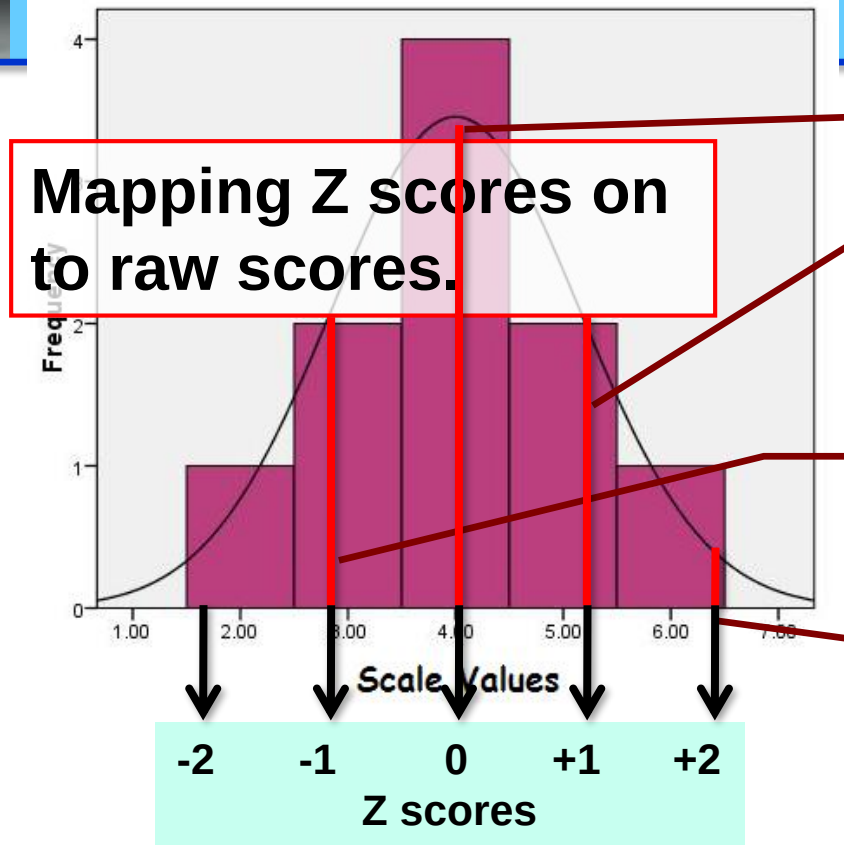


In this distribution M = 4 and one standard deviation [S] = 1.14.

Standard deviations represent variance both above and <u>below</u> the M

About 34% of cases are between the M and one standard deviation above the mean, or between 4 → 5.14.

Another 34% are between M and 1 standard deviation below the mean…4 → 2.86



M = 4 (Z = 0)

S = 1.14

Z of +1 = M + 1S
= 4 + 1.14 = 5.14

Z of -1 = M - 1S
= 4 - 1.14 = 2.86

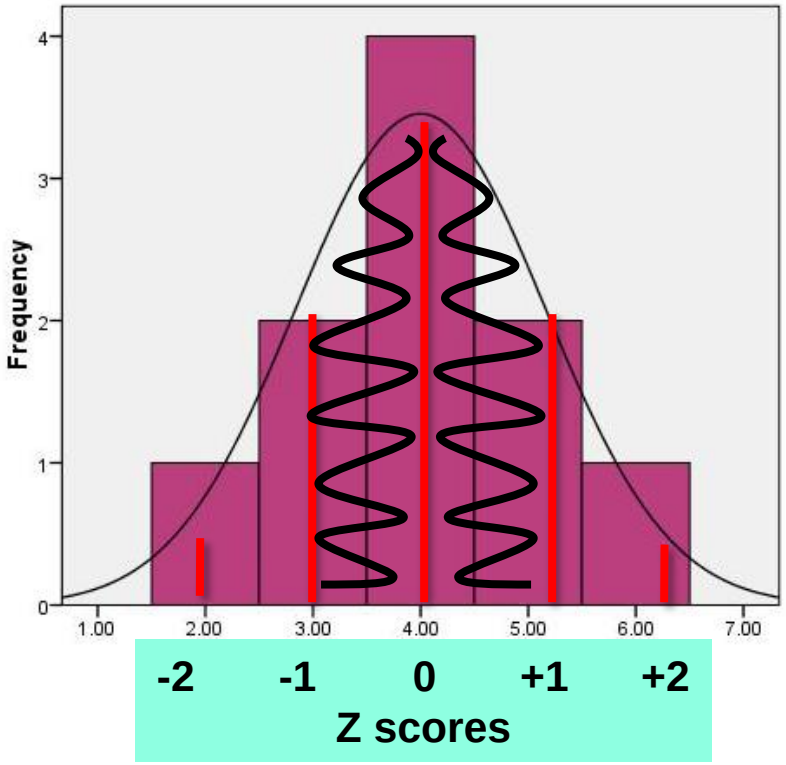
Z of +2 = M + 2S
= 4 + 2.28 = 6.28

Z scores translate raw scale values into **standard deviation units**.

The Z scores show what a much larger, hypothetical distribution would look like with M = 4 and S = 1.14.

This becomes the basis for inferential statistics using these data.

Transforming raw scores to Z scores



- **The *M* of the distribution has Z = 0**
- **Each Standard deviation unit** (S = 1.14 in this distribution) **is a Z of 1.**
  - **About 34% of cases are between:**
    - ✓ **M → 1 standard deviation above the mean**

      Z = 0 to Z = +1; 4 → 5.14 in raw scores.

    - ✓ **M → 1 standard deviation *below* the mean**

      Z = 0 to Z = -1; 4 → 2.86 in raw scores.

A distribution of scores can be segmented into…?

A. Standard Deviation units.
B. Z scores
C. Sums of squares
D. Degrees of freedom
E. Variance

A distribution of scores can be segmented into…?

A. Standard Deviation units.
B. Sums of squares
C. Z scores
D. Degrees of freedom
E. Variance

- ✓ Each unit of Z represents one Standard Deviation.
- ✓ A score one standard deviation above the Mean has Z = 1.
- ✓ Z units or Standard Deviation units reflect the % of scores below (or above) the score in question.

X – M ….?

A. How far a score is from the Mean
B. How much variance there really is in the sample
C. Distance of a score from M adjusted by n
D. Distance of a score from M adjusted by S

X - M ….?

A. How far a score is from the Mean

B. How much variance there really is in the sample

C. Distance of a score from M adjusted by n

D. Distance of a score from M adjusted by S

Z tells us…

A. How far a score is from the Mean
B. How much variance there really is in the sample
C. Distance of a score from M adjusted by n
D. Distance of a score from M adjusted by S

Z tells us…

A. How far a score is from the Mean

B. How much variance there really is in the sample

C. Distance of a score from M adjusted by n

D. Distance of a score from M adjusted by S

- ✓ Z calibrates not only how far a score is from the Mean, but the variance of other scores above or below the M.
- ✓ That variance is represented by the Standard Deviation of the scores [S].
- ✓ This tells us how much one score deviates from M relative to how much other scores deviate from M.

Both the range and the standard deviation are examples of this…

A. Mean
B. Ratio scale
C. Degrees of Freedom
D. Sum of Squares
E. Variance

Both the range and the standard deviation are examples of this…

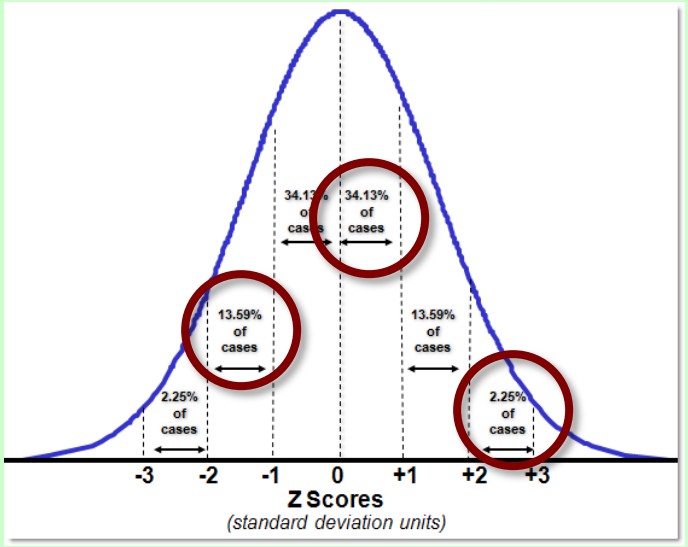
A. Mean
B. Ratio scale
C. Degrees of Freedom
D. Sum of Squares
E. Variance

"Variance" has two meanings in statistics:

- ✓ **The general concept of scores differing from each other in a sample**
- ✓ **A statistical formula:**
  - **Distance from the highest to lowest score (range).**
  - **Amount the scores vary around the Mean (Standard Deviation).**

Summary

- **Standard deviation is the basic metric of variance in a sample.**
- **Each standard deviation above or below the Mean represents a fixed ("standard") % of cases.**
- **Z tells us the number of standard deviation units a score is above or below the mean.**



- **Z =** $\dfrac{\text{Distance of a score from the Mean (X – M)}}{\text{Standard Deviation of all scores in the distribution (S)}}$

- **A score right at the *M* has Z = 0.**
- **Each standard deviation a score is from M = Z score of 1**
- **Z can tell us the % of scores above or below any given score.**

Next module

In the next module we will discuss how we use Z scores to evaluate data

Shutterstock.com